科技查新中检索词智能抽取系统的设计与实现*

王培霞[1,2]　余　海[1,2]　陈　力[1,2]　王永吉[1]
[1](中国科学院软件研究所　北京 100190)
[2](中国科学院大学　北京 100049)

摘要: 【目的】解决科技查新领域检索词选择时的主观性强、手工工作量大、不规范、费时费力的问题。【应用背景】为了实现检索词抽取过程的自动化、智能化、规范化,本文提出利用科技查新过程检出的实时相关语料作为领域知识的来源,并对语料组成类型与关键词抽取效果之间的关系进行讨论。【方法】通过关键词抽取、领域特征扩展相结合的递进式迭代抽取方式实现科技查新领域检索词的智能抽取。【结果】通过与实际查新案例所采用的检索词对比,发现使用本方法两次迭代后抽取 10 个检索词,召回率达到 80%。【结论】基于查新过程中检出文献构成的动态相关语料进行检索词的迭代抽取有助于快速、准确锁定绝大部分检索词,提高检索的效率和效果。

关键词: 科技查新　检索词　关键词抽取　网络爬虫
分类号: TP391

1 引　言

科技查新是一种通过文献检索、对比和分析,查证查新项目新颖性的信息咨询服务工作。根据查新时的检索范围,可分为国内查新、国外查新和国际查新三类。其中,文献检索是科技查新的基础,涉及检索式的组配,其中检索词的选定对检索结果起着关键性的作用,是影响科技查新质量的三个重要因素之一[1],尤其是在进行国际查新时,英文检索词的正确与否会直接影响到检索结果的新颖性、全面性和准确性(简称三性)。

检索词是指表征查新项目主题内容的、具有实质意义的词语,对揭示和描述查新项目主题内容是重要的、关键性的词语。目前,国内科技查新咨询机构在开展科技查新业务时,为了保证查新结果的三性,通常使用主题词和关键词作为检索词进行文献检索。

主题词也被称作叙词和受控词,是规范化的检索语言,是对某一概念的同义词、近义词进行规范化处理后确定的检索词[2],主要来自于规范化词表。然而,文献数据库存在标引不规范的问题,而且查新员在规范化的过程中需要借助专业的叙词表或受控词表,这些规范化词表通常因维护周期长而比较固定,对一些体现新的技术、方法、理论的词不能及时登录,科技查新业务的文献检索则需要体现技术、方法、理论的创新性,因而使用主题词作为检索词进行检索容易造成漏检;而关键词(也称自由词),它基于文献的标题、摘要、关键词甚至全文进行词的索引、检索,自由度高,不需考虑标引、规范性、主题词表的更新时效等问题,是电子信息资源的重要检索途径,本研究以关键词检索作为研究重点。

通讯作者: 王永吉, ORCID: 0000-0001-7472-7489, E-mail: ywang@itechs.iscas.ac.cn。

*本文系国家自然科学基金项目"云计算环境下的隐蔽信道机理研究"(项目编号: 61170072)、国家自然科学基金青年科学基金项目"移动智能终端隐蔽信道机理研究"(项目编号: 61303057)和中国科学院、国家外国专家局创新团队国际合作项目"安全攸关软件理论和构造方法"的研究成果之一。

当前，查新人员通常需要手工完成检索词的发现、筛选、补充、扩展及最终选定，依据委托人提供的科技查新项目资料，以委托人提供的关键词为参考，结合科学技术要点、查新点及其他补充材料初步找出符合查新主题的关键词[3-4]，必要时还需要进行关键词扩展，利用专业词表、辞海、词典、术语标准等工具书及已检出的文献获取检索词的规范名称、同义词等进行关键词扩展。此外，查新人员在检索过程中会进行试检，根据检出文献的相关信息进一步判断检索词是否合适，从而对检索词进行调整，为了获得满意的检索效果，此过程通常会反复多次。

由此可见，检索词的选择要经过相关文献的检索、检出文献的浏览、分析、综合、调整的迭代过程，需要向多个文献数据库多次提交检索请求，反复试检，在每次试检后还需要对检出文献进行检索词分析、调整，经过多次循环迭代后才能最终确定检索词，因而存在工作量大、费时费力，对查新人员的耐性也是一个不小的考验。另外，此过程与查新人员的专业水平、经验、知识结构等关系密切，具有较强的主观臆断性，难以规范，因而也会直接影响最终的检索效果，从而会影响到科技查新报告的质量。

为了克服检索词选择主观性强、手工工作量大、不规范、费时费力的问题，实现检索词选择的科学性、规范性，本文引入与科技查新项目有关的动态文献语料，同时立足查新项目信息，以实时获取的与查新项目相关的动态语料作为领域知识的来源，采取关键词抽取、领域特征扩展相结合的递进式迭代方式进行检索词的抽取。

2 相关工作

2.1 关键词自动抽取

关键词自动抽取技术在文献检索、自动文摘、文本聚类、分类等领域都有很广泛的应用，如信息检索领域，好的关键词可作为全文索引的补充，有助于用户发现相关文档。

关键词的自动抽取过程通常分为两步[5]:

(1) 候选关键词识别: 使用某些启发式规则(去掉停用词; 词性选择上只保留名词、形容词、动词; 使用外部资源如维基百科; N 元语法等)抽取出词或词组作为候选关键词;

(2) 候选关键词选择: 使用基于监督或非监督的方法判断哪些候选关键词是正确的。

大量的关键词抽取研究工作集中在候选关键词的选择上，主要分为有监督和无监督两类。早期的基于监督的方法将关键词抽取看作一个分类问题，通过利用已标注语料进行数据训练，构建学习模型，进而判断词语是否属于关键词类别，学习算法包括朴素贝叶斯[6]、决策树[7]、最大熵[8]、多层感知器[9]、向量空间模型[10-11]等。基于监督的方法需要事先用作者提供的关键词或专家标注的关键词进行语料标注，因而成本比较高。另外，分类器在判定一个词是否为关键词的过程中独立于其他词，而 Turney[12]的研究显示，关键词的选择并不是相互独立的，即之前选择的关键词会对后面的关键词选择有影响。

关键词抽取技术主要分为三类: 基于统计特征的关键词抽取、基于主题模型的关键词抽取和基于词图模型的关键词自动抽取方法。

基于统计特征的关键词抽取通过计算词的某些特征(如词频、N-gram[13]、TF-IDF 值、信息熵等)，结合其位置标记(如题名、段首、首次出现的位置等)为词分配权重，根据权重大小顺序提取关键词。如 Frank 等[6]在构造模型时使用 TF-IDF 得分、关键词第一次出现的位置两项特征; 潘丽敏等[14]在关键词抽取过程中除了使用 TF-IDF 得分，还融合了关键词短语的长度、短语是否在题名中出现、短语在文档中的分布情况、最大词频和最小词频等特征。Hulth [15]加入了语言学知识如名词短语块(NP Chunking)和词性标注(Part-of-Speech tags , POS)，使抽取正确率大为提高。

基于主题模型的关键词抽取方法以基于 LDA 的关键词抽取方法应用最为广泛[16-18]，通过大量已知的“词语–文档”矩阵和一系列训练推理出“文档–主题”分布和“主题–词语”分布，该方法认为在文档中主要主题中的主要词语更有可能被识别为关键词。主题模型需要对数据进行训练得到，关键词抽取的效果与训练数据的主题分布关系密切。

基于图模型的关键词抽取以 TextRank 算法[19]为典型代表，其提出受到 Google 的 PageRank[20]思想的启发。它基于文档构建一个词图，图中每个节点对应一个候选关键词，每条边代表候选关键词之间的关系，当两个词语在一个观测窗口出现，那么它们之间就建

立了关联关系。对每个节点，与其相连的每一条边都认为是一次“投票”，其重要性由与其相连的其他节点决定，通过循环迭代计算，按词语的重要性得分高低最终确定关键词，从而实现了基于单文档的关键词抽取。

另外，关键词抽取时除了使用文档本身的特征以外，还引入了各种领域知识。主要分为两类:

(1) 词典类。如叙词表[21-24]、互联网词典[14]、术语数据库[9]或维基百科[9]等通用领域词典、百科资源。维基百科含有丰富的百科词条，每个维基百科词条可看作是一个独立的概念。其中，文献[9]利用一个词为维基百科词条的可能性作为关键词特性的判断依据。

(2) 语料类。主要包括领域相关语料库、通用语料库、对比语料库。已有实验表明，相同领域的文档对关键词的抽取效果有很大的帮助: 文献[25]使用与待抽取文档相同领域的语料信息，即作者标注关键词的长度、成分和词频信息; 文献[6]在模型训练时同样使用与待抽取文档相同领域的文档，且抽取效果与所需文档的数量成正相关，这表明，相同领域的文档越多抽取效果越好。但这也带来一个现实的问题，领域有关的语料均来自于人工获取，且数量庞大，势必要耗费大量时间和人力成本，而且这些语料都是静态的语料，会随着时间的变化而变得过时，时效性不强。在基于科技文献的关键词抽取领域，文献[26]提出将文献的标题和关键词作为种子词语，基于开放领域的语料库利用 Word2Vec 找出相似的词语作为候选词，实际上，开放领域的语料库本身对于检索词而言在领域相关性方面并不占优势。Lopes 等[27]在计算领域相关性时使用了对比语料库，但是对比语料库语料选择以及其语料库的大小都会直接影响到抽取效果，同时，在实际应用中，语料库的人工获取也是一个严峻的问题。

2.2 检索词推荐

与本研究相关的另外一种就是基于数据挖掘的检索词推荐技术。通常，推荐分为基于规则过滤、基于内容过滤、基于协作过滤以及多方法混合的推荐方法。此类推荐系统通常基于用户使用过程中产生的历史行为记录，如检索日志等，通过对用户行为进行建模，挖掘其中的行为规律，因而检索词的推荐技术建立在已有系统的有效运行基础之上，重在已有静态数据的挖掘，而本研究重在检出相关文献的检索词动态抽取，因而侧重点不同。

2.3 本研究主要贡献

综上，本研究是基于查新项目的题目、关键词、科技要点、查新点及检索过程中产生的文献的相关知识进行检索词的抽取，自动抽取与输入的检索词有关的领域特征词作为候选检索词。本研究具有以下几个特点:

(1) 基于查新项目及与查新项目有关的动态文献语料进行检索词抽取。而传统的关键词提取主要面向科技论文、论文摘要、网页等单文档或使用语料库，本研究的主要抽取对象则为科技查新项目申请及检索过程中产生的文献语料，是由科技查新申请及文献检索阶段产生的多个甚至很多个相关文献组成，这些文献具有领域相关性、数据量大、内容丰富、内容权威性强等特点，通过网络爬虫在线获取，可以与数据源保持同步，具有动态性、实时性，不会随着时间的变化而发生过时。

(2) 侧重于领域知识的引入。传统的关键词抽取技术以文献标引为目的，因而在关键词抽取过程中仅限于文献的题目、摘要、正文内容，除了利用检出文献的题目、摘要以外，还利用文献的关键词，而这些关键词通常是由作者选定的，是表示领域概念的基本要素，具有较强的指示性、领域区分能力，是检索词的重要来源。

(3) 本研究抽取的候选检索词主要辅助查新员快速找出相关检索词，充分利用科技文献本身作者标注关键词的领域专业特性，有助于防止漏检、提高国际查新的查全率、查准率。这与以文献标引为目的的关键词抽取有很大区别。

(4) 候选检索词的抽取过程具有动态性、递进性。传统的关键词抽取一次即可实现关键词的抽取，而面向科技查新的检索词抽取过程具有交互性和动态性。查新员在检索过程中通过迭代、反复检索，逐步递进地调整检索词以获取满意的检索效果。

基于以上 4 点特性，本文首先对文献中作者标注的关键词在题名、摘要的分布情况做抽样统计，分析科技文献的关键词与题名、摘要的分布关系，然后基于检出的语料按题名、关键词、摘要三种类型以词语共现关系作为词语之间的关联关系，以共现词语的词频(Term Frequency)作为关联关系的强度构建词图并

进行候选检索词的抽取实验，并通过对比实验结果，分析三者在检索词抽取效果方面的贡献程度。最后，通过关键词抽取、领域特征扩展相结合的递进式迭代抽取方式实现科技查新领域检索词的智能抽取，并通过实际的查新案例进行说明。

3 系统设计

面向科技查新领域的检索词智能抽取系统由基于网络爬虫的文献在线检索、检索词智能抽取两部分组成。由于检索词的智能抽取与查新项目有关，建立在与查新项目相关的动态语料的基础上，因而语料的获取是抽取系统的重要组成部分。

系统采用 Spring Web MVC 和 Hibernate 框架开发，前者是一种基于 Java 的实现了 Web MVC (Model-View-Controller) 设计模式的请求驱动类型的轻量级 Web 框架，后者是一个开放源代码的对象关系映射框架，它对 JDBC 进行了非常轻量级的对象封装，本地数据采用 MySQL 数据库存储采集到的文献信息，实现了数据、业务与展现的分离。

3.1 语料获取

语料获取主要通过网络爬虫在线获取各个文献检索系统检出的科技文献信息，其流程如图 1 所示：

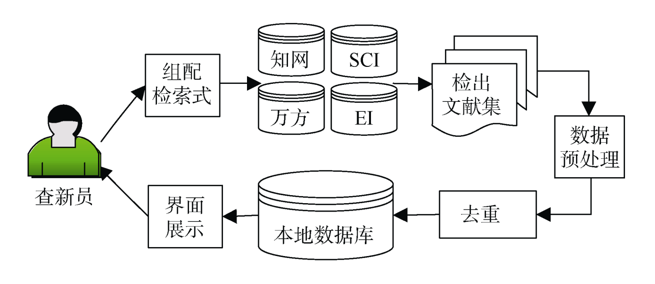


图 1 文献在线检索流程

网络爬虫是一个抓取网页内容的程序，利用网页格式特征进行网页分析[28]。在本系统中，数据预处理主要利用网页的标签结构分析出文献的标题、摘要、中英文关键字、作者等信息并存储到本地数据库服务器。本系统还设置了搜索日志，解决热门检索词的因反复检索造成的时间耗费较长的问题，对于重复的检索式可以将以前抓取的结果展示给用户，另外，系统还设置了去重处理，防止数据的重复采集。

3.2 检索词智能抽取

检索词智能抽取的目的是基于科技查新文献检索获取的动态语料生成可供查新员直接使用的检索词，其流程如图 2 所示：

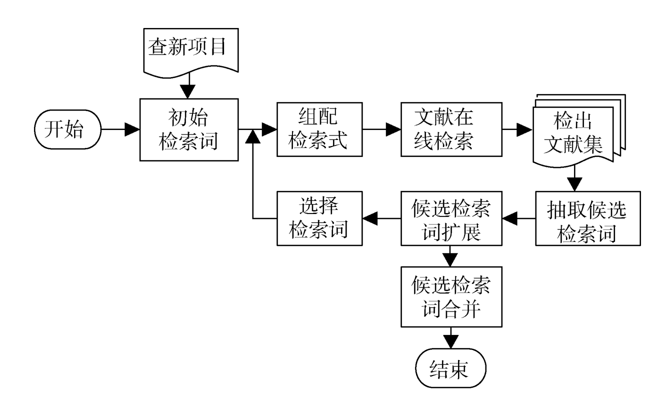


图 2 检索词智能抽取流程

检索词智能抽取过程由以下几个步骤组成：

(1) 语料获取。根据查新项目的题目自动生成检索词并组配检索式进行跨库检索，实时获取检出文献信息，存储到本地数据库中。

(2) 候选检索词抽取。基于步骤(1)生成的动态语料自动抽取 10 个候选检索词。

(3) 候选检索词扩展。对步骤(2)生成的候选检索词进行领域特征扩展，生成检索词表。

(4) 判断。人工核查检索词表是否满足需求，如不满足，可以选择合适的检索词手动组配检索式重复步骤(1)到步骤(3)。

(5) 合并。根据检索词表中每个词的重要性进行合并，生成最终的检索词列表。

4 检索词抽取方法

通常情况下，重要的术语在相同领域的科技文献语料中出现的概率较高。而且，科技文献具有丰富的文档结构，通常包含标题、摘要、作者标注的关键词等相关信息，在语言表达上具有领域专业性。另外，检索词与科技查新项目密切相关，而且很多都是专业术语。

基于以上考虑，本文提出利用科技查新过程中检出语料与查新项目的领域相关性，在首次获取相关语料时利用标题中的有关词组配检索式获取领域相关语料。基于检出语料，抽取出具有提示特征的关键词，在此基础上进行领域特性的扩展，生成相关的候选检索词。该方法基于这样一个假设：查新项目的标题是查新主题的简明描述，查新项目的科技要点是查新项目

的详细描述。

考虑到查新项目的标题在主题表达上并不能反映查新项目的所有领域特征，在这里，使用多次迭代、反复检索的方式，在后续的每一次迭代过程中，都由查新员从已生成的检索词表中选择检索词或检索词的组合组配检索式进行领域相关语料的获取。

检索词的抽取过程主要包括三个步骤：基于检出文献语料抽取候选检索词；对候选检索词进行领域特性的扩展；合并。

4.1 抽取候选检索词

(1) 语料分析

候选检索词的抽取语料来源于各个文献数据库根据接收到的检索式返回的文献信息，通过爬虫程序抓取文献的标题、关键词、摘要等信息作为候选检索词的抽取对象。

科技文献的关键词是为了文献标引工作而从学术论文中或之外选择出用以表示全文主题内容信息款目的单词和术语，是未规范的自然语词[29]。在计量统计学领域，研究者认为文献的关键词是表示领域概念的基本要素，它主要用来从宏观上研究领域知识的结构特征或者从微观上使用一些"重要"词分析一个领域的主要研究主题的细节及其它们之间的关系[30]。因此，科技文献的关键词本质上反映了该领域的知识结构和主题特性，是领域特征词，在某一领域中具有较强指示性、领域区分能力，因而是获取高质量检索词的重要来源。

为了更清楚地观察科技文献中作者标注关键词在标题、摘要的分布情况，本文基于知网数据库和万方数据库所收录的文献做抽样统计，结果如图 3 所示：

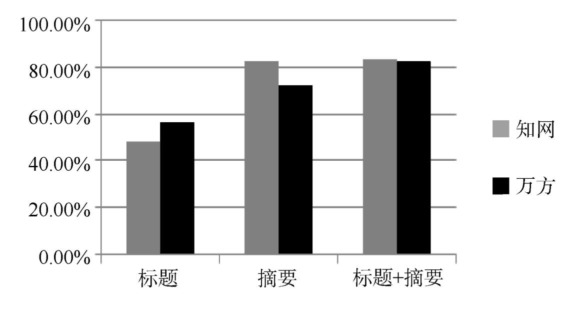


图 3 论文题名、摘要与关键词的分布比率分析

在这里，使用平均比率 avg(t) 表示关键词在标题、摘要及标题和摘要的组合中的分布比例。

$$\mathrm{avg}(t)=\frac{n(t_i)}{|t|} \tag{1}$$

其中，$n(t_i)$ 表示文献的关键词在题名、摘要或二者的组合文本中出现的个数，$|t|$ 表示文献关键词的总个数。

观察图 3 发现，关键词在文献标题的分布约占 50%，而论文摘要中包含的关键词数量更多，尤其是知网，平均比率达到 80%以上，标题和摘要二者的组合对关键词的包含率均达到 80%以上，因而在进行候选检索词的选择时，文献本身提供的关键词是候选检索词非常重要的来源。

为了更加清楚地查看文献的标题、关键词、摘要在检索词抽取时所产生的不同效果，分别基于文献的标题、关键词、摘要及其它们之间的组合文本进行不同方法的检索词抽取实验。

(2) 基于改进的 TextRank 候选检索词抽取

经典的 TextRank 算法在进行关键抽取时，一个文档用一个无向图来表示，图的节点表示词，在给定的文本窗口内任意两个词之间都构建一条边，对图中的任一节点来说，其重要性得分由其相邻节点的贡献组成，基于 PageRank 方法进行词语重要性的迭代计算，在该算法中，任意两个相邻词之间的关联度是相同的，未探讨相邻词语之间的影响力强弱。然而在科技查新的文献检索过程中，检出文献通常数量较多，且文献之间以检索词为纽带，互相具有一定的关联性，词频较大的词或词组在一定程度上反映了检出文献的主题信息倾向，因而在本文中，以词频为基础对经典的 TextRank 算法进行改进(记为 MF_TR)，以词出现的频次作为词语的重要性影响因子，重要性转移矩阵同样以词频为基础，候选词的重要性得分计算公式如下：

$$\mathrm{Score}(i)=(1-d)+d\times\sum_{j\in set(i)}\frac{\mathrm{tf}(i)}{\sum_{k\in set(j)}\mathrm{tf}(k)}\mathrm{score}(t_j) \tag{2}$$

其中，set(i)为词 i 的共现词，tf(i)为词 i 的词频。

通过公式(2)计算每个词的重要性得分，按得分大小进行降序排序，选择排序靠前即重要性较高的前 10 个词作为候选检索词。

4.2 候选检索词的扩展

对获得的关键词基于检出文献的关键词或摘要等相关文本语料进行领域特征的扩展。由于大部分中文词性标注系统基本上以新闻语料进行训练，而科技文

献的关键词却很少出现在新闻语料中，另外，分词系统在分词的过程中一些专业术语也会因各种原因被分割开来，如词“全沟硬蜱”，经过分词后，“全沟”被单独作为一个词对待，因而为了对这类错误进行纠正，笔者对每个抽取出来的关键词在检出文献的关键词集合中查找是否有包含该关键词的领域特征关键词，如果存在，那么作为候选检索词，并计算其作为重要性得分。考虑到摘要是正文主题内容的概括性描述，因而使用候选检索词在摘要文本中出现的频次作为重要性调节因子，并结合候选检索词的短语特性(公式(3))计算其重要性得分(公式(4))。

$$\mathrm{GDC(T)}=\frac{|\mathrm{T}|\log_2 \mathrm{freq(T)freq(T)}}{\sum_{t\in \mathrm{T}} \mathrm{freq}(t)\times \mathrm{N}} \quad (3)$$

$$\mathrm{DGDC(T)}=\mathrm{tf(T)}\times \mathrm{GDC(T)} \quad (4)$$

其中，T 为候选检索词，|T|为其所包含的词语 t 的个数，而不是所含的字的个数，N 为检出文献的数量，tf(T) 为待抽取文档描述性文本所包含的候选检索词的频次，如果是科技文献，则为摘要或正文中 T 出现的次数，本实验中使用科技要点作为科技查新项目的描述。对于 T 出现次数为 0 的情况，给定一个特定的初始值 0.1F，这意味着如果 T 在相关语料中即使作为领域关键词的短语特性具有很高的值，但是如果在描述性文本中没有出现，那么其最终的领域重要性得分将被拉低，事实上，tf(T)可看做一个权重因子，对候选关键词 T 的最终领域重要性得分起着调节作用。

4.3　检索词合并

由于在实际的查新过程中，检出文献对检索词有很重要的影响，因此以上三个步骤可重复进行，除了第一次自动生成检索式以外，后续均基于生成的检索词表或查新项目的关键词等信息手工组配检索式，最后对扩展后的检索词集合根据重要性进行排序，取重要性最高的前 10 个作为最终的检索词。

5　实　验

5.1　数据源

本实验采用真实的科技查新委托申请数据，其题目、关键词、科技要点、查新点及其他信息均来自于客户的填写。用来提取特征词的科技文献源于科技查新业务平台采集的相关文献。用户通过初始的检索词组配检索式，并将检索请求发送给各个文献检索库，利用开源网页分析工具 HtmlParser 对页面进行分析，提取文献题名、关键词、作者、期刊名称及期刊号、摘要、受控词、非受控词等文献相关信息，并存储在本地数据库。目前，科技查新业务平台支持中英文检索，中文包括万方、知网，英文包括 Web of Science、EI 数据库。

5.2　数据预处理

数据预处理过程包括数据格式规范化、数据去重和文本预处理。

(1) 数据格式规范化：由于不同的文献数据库在文献收录的过程中采用不同的标准，因而有必要对采集的文献格式进行统一、标准化处理。

(2) 数据去重：由于中国知网、万方数据、重庆维普等中文数据库存在重复收录文献的情况，如维普数据库对万方期刊数据库的重复率为 93.6%，对中国知网的重复率为 94.1%[31]，因而有必要在数据预处理阶段进行去重。

通过对万方、知网等采集的数据初步分析，将文献资源的重复分为两类：

①数据库重复收录。即同一文献被一个或多个数据库重复收录两次或以上。

②重复出版。一个研究成果被发表在不同的杂志或出版物上。

去重算法利用论文题名、第一作者、期刊和年份作为去重的依据，并对所检测的重复进行分类，以便后续针对不同类型的重复采取不同的处理策略。对于第一种重复类型，需要删除重复的文献。对于第二种重复类型，暂时保留。

(3) 文本预处理：此过程主要从检出的文献识别候选关键词。

分词、词性标注、词性选择：利用 HanLP 工具包提供的词性标注器(Part-of-Speech tagging)为检出文献进行分词并标注词性，除掉标点符号、数词、区别词、连词、叹词、拟声词、介词、量词、助词、语气词、状态词、代词，其他类型词性的词被保留。

去掉停用词：针对文献的摘要特点，在通用停用词表的基础上添加科技文献的摘要中常用的词，如研究、具有、采用、进行、结果表明、应用、方法、问题、分析等。

5.3 候选检索词抽取实验

(1) 相关方法介绍

本文选择 4 种常见的关键词抽取方法作为对比，包括 Most Frequent [32](MF)、Term Frequency - Inverse Document Frequency (TF-IDF)、Latent Dirichlet Allocation (LDA)、TextRank(TR)，选择重要性得分较高的 10 个词作为候选检索词。

①MF：自动抽取关键词的最简单的方法就是考虑出现频率最大的词。在本文中，每个词语的重要性得分由检出文献中该词的出现频次决定，即：

$$score(t_i) = f(t_i) \tag{5}$$

②TR：基于文献[19]对经典 TextRank 的描述，词语的重要性得分为：

$$score(t_i) = (1-d) + d \times \sum_{e_{t_i,t_j} \in \varepsilon} \frac{1}{\sum_{e_{t_j,t_k} \in \varepsilon} \delta_{t_j,t_k}} score(t_j) \tag{6}$$

其中，$d \in [0,1]$ 为阻尼系数(Damping Factor)，在实验中，取值为 0.85f，e_{t_i,t_j} 为 t_i 和 t_j 的边，δ_{t_j,t_k} 代表词语是否共现，若有共现关系，其值为 1，反之为 0。

③TF-IDF：其基本思想为一个词语的重要性由该词的词频和它在语料中分布决定。文献中词汇的频度越高，在某种程度上就意味着其与文献所表达主题相关的可能性较高，但是如果该词汇在大量句子中都出现，则其重要性会因为特征不明显而降低。本文采用如下公式计算词语的重要性得分：

$$score(t_i) = TF\text{-}IDF(t_i) = f(t_i) \times \log \frac{N}{n(t_i)} \tag{7}$$

其中，$f(t_i)$ 是 t_i 在检出文献中出现的总次数，N 为检出的文献总个数，$n(t_i)$ 为包含词语 t_i 的文献个数。

④LDA：基于文献对 LDA 模型的描述，令 n 为语料库中文档的个数，φ 表示 LDA 中主题–词语的概率分布，φ_w^z 表示词语 w 在主题 z 中的概率，θ 表示文档–主题的概率分布，$\theta_{z=j}^{d_i}$ 表示文档 d_i 中主题 z 的概率。本实验中，一个词语 w 在文档集 D 中面向 t 个主题的概率通过以下公式计算：

$$P(w|D) = \sum_{i=1}^{n} \sum_{j=1}^{t} \varphi_w^{z=j} \theta_{z=j}^{(d_i)} \tag{8}$$

数值的大小可以反映词语在文档集中面向主题的重要性，按照值的大小顺序选择前 σ 个词作为关键词，在本实验中 σ 取值为 10。

(2) 实验及方法对比

①语料与抽取效果

首先，以“石墨烯在锂电池中的应用及发展前景”为例进行抽取。

在实施文献检索阶段，为了控制在线检索的时间、爬虫获取及解析数据的时间以及返回过量无用文献，以每个文献数据库在线获取不超过 110 条记录为限，采取递进式、交互式的方式对获取的文献进行实验。

使用检索式“石墨烯”进行检索，在万方检出 110 条记录，基于检出文献的题名、关键词、摘要以 MF_TR 检索词提取方法(其中，文本共现的观测窗口 w 默认为 5，迭代运算的终止条件为迭代次数大于 200 或两次迭代结果的差异值小于等于 0.001f)分别提取 10 个候选检索词，其结果如表 1 所示：

表 1 检索式为“石墨烯”候选检索词抽取实验结果

序号	关键词	题名	摘要
1	石墨烯	石墨烯	石墨烯
2	graphene	graphene	graphene
3	氧化	制备*	氧化
4	oxide	氧化	制备*
5	还原	性能*	结构
6	石墨	preparation*	还原
7	结构	oxide	性能*
8	纳米	复合材料	复合材料
9	复合材料	还原	材料*
10	化学	synthesis*	rgo*

(注：以基于关键词抽取的候选检索词为基准，标注*的候选词为新出现的词。)

以检索式“锂电池”候选检索词抽取实验结果，其结果如表 2 所示：

表 2 检索式为“锂电池”候选检索词抽取实验结果

序号	关键词	题名	摘要
1	锂电池	锂电池	锂电池
2	battery	battery	电池
3	lithium	lithium	性能
4	电池	batteries*	充放电*
5	储能	li-ion*	系统*
6	磷酸铁	energy*	循环*
7	soc	性能	battery
8	状态	based*	模型*
9	性能	储能	soc
10	材料	storage*	容量*

观察表 1 和表 2 可知：基于检出文献的题名、关键词、摘要进行候选检索词的抽取实验中，检索式“石墨烯”共有 16 个不同候选检索词，其中共同出现的词有 5 个，占比 31.25%；检索式“锂电池”共有 20 个不同候选检索词，其中共同出现的词有 3 个，占比 15%。由此说明，对不同的检索式，所抽取的候选检索词在题名、关键词和摘要中共现的分布差别较大。在抽取的效果方面，基于关键词、题名抽取的候选检索词在同位、上下位关系方面要好于基于摘要的抽取方式。基于检出文献关键词语料抽取的候选检索词领域专业

性、全面性方面要好于其他两种方法，因而关键词是抽取候选检索词的重要来源。

为了验证关键词与其他项(题名、摘要)的组合效果是否能提升检索词的抽取质量，以检索式"石墨烯"为例，以基于检出文献的关键词为基础，分别与题名、摘要及二者共同组合以 MF_TR 检索词提取方法分别提取 10 个候选检索词，其结果如表 3 所示：

表 3　基于检出文献的题名、关键词、摘要组合的检索词抽取实验结果

序号	关键词	关键词+题名	关键词+摘要	关键词+题名+摘要
1	石墨烯	石墨烯	石墨烯	石墨烯
2	graphene	graphene	graphene	graphene
3	氧化	氧化	氧化	氧化
4	oxide	oxide	制备*	制备*
5	还原	制备*	结构	性能*
6	石墨	还原	还原	还原
7	结构	性能*	性能*	结构
8	纳米	复合材料	oxide	oxide
9	复合材料	preparation*	复合材料	复合材料
10	化学	结构	材料*	材料*

观察表 3 可知：以关键词为基准，当分别与题名、摘要及二者共同组合进行候选检索词的抽取时，分别有三个新的候选检索词进入观测表格。关键词、摘要的组合抽取效果与关键词、摘要、题名三者的组合抽取效果相当，只是候选检索词的顺序稍微有些变化。关键词与题名的组合抽取效果，制备、性能、preparation 三词替换了只以关键词抽取的石墨、结构、纳米三词实际上效果并未得到明显改善。

为了验证是否因为检出文献太少导致这样的结果，以"石墨烯"为检索式，对检出文献的数量增加到 231 条，实验结果如表 4 所示：

表 4　基于检出文献的题名、关键词、摘要的检索词抽取实验结果

序号	关键词(110)	关键词(231)	题名(110)	题名(231)	摘要(110)	摘要(231)
1	石墨烯	石墨烯	石墨烯	石墨烯	石墨烯	石墨烯
2	graphene	graphene	graphene	graphene	graphene	graphene
3	氧化	氧化	制备	制备	氧化	氧化
4	oxide	oxide	氧化	preparation	制备	制备
5	还原	还原	性能	性能	结构	结构
6	石墨	石墨	preparation	氧化	还原	性能
7	结构	纳米	oxide	oxide	性能	还原
8	纳米	复合材料	复合材料	复合材料	复合材料	复合材料
9	复合材料	结构	还原	还原	材料	材料
10	化学*	性能*	synthesis	synthesis	rgo*	纳米*

(注：标*的词表示有变化的词语。)

观察表 4 可得出，在观测窗口(前 10 个词)内，检出文献数量的增多只对抽取词语有轻微影响，词语的变化限制在一个词语范围内，即实际上对最后的抽取效果影响甚微，反而占用更多的资源，增加了时间成本。

②相关方法对比

使用检索式"石墨烯 and 锂电池"进行检索，在万方数据库检出 131 条文献信息后基于题名、摘要、关键词分别基于最大词频法、TR 方法、TF-IDF 方法及 LDA 的方法提取 10 个候选检索词，其结果如表 5 至表 9 所示：

表 5　基于题名的检索词抽取方法对比

序号	MF	MF_TR	TR	TF-IDF	LDA
1	石墨烯	石墨烯	石墨烯	石墨烯	石墨烯
2	性能	性能	性能	制备	性能
3	制备	制备	制备	性能	制备
4	材料	材料	材料	材料	材料
5	锂电池	锂电池	锂电池	锂电池	锂电池
6	纳米	纳米	电池	纳米	纳米
7	复合材料	复合材料	合成	复合材料	复合材料
8	电化学	电化学	复合材料	电化学	电化学
9	合成	合成	纳米	合成	合成
10	电池	电池	电化学	电池	电池

表 6　基于摘要的检索词抽取方法对比

序号	MF	MF_TR	TR	TF-IDF	LDA
1	石墨烯	石墨烯	石墨烯	纳米	石墨烯
2	材料	材料	材料	石墨烯	材料
3	性能	性能	性能	复合材料	性能
4	纳米	纳米	制备	容量	纳米
5	电池	电池	电池	电池	容量
6	容量	制备	结构	材料	电池
7	制备	结构	纳米	循环	制备
8	结构	容量	复合材料	性能	结构
9	循环	循环	容量	制备	循环
10	复合材料	复合材料	循环	mAh	复合材料

表 7　基于题名、摘要的检索词抽取方法对比

序号	MF	MF_TR	TR	TF-IDF	LDA
1	石墨烯	石墨烯	石墨烯	纳米	石墨烯
2	材料	材料	材料	复合材料	材料
3	性能	性能	性能	容量	性能
4	纳米	纳米	制备	电池	纳米
5	制备	制备	电池	循环	制备
6	电池	电池	纳米	mAh	电池
7	容量	结构	结构	Li	容量
8	结构	容量	复合材料	石墨烯	结构
9	循环	复合材料	容量	g-	复合材料
10	复合材料	循环	循环	Fe	循环

表 8　基于关键词的检索词抽取方法对比

序号	MF	MF_TR	TR	TF-IDF	LDA
1	石墨烯	石墨烯	石墨烯	石墨烯	石墨烯
2	材料	电池	电池	材料	材料
3	电池	材料	graphene	电池	电池
4	锂电池	锂电池	锂电池	锂电池	锂电池
5	锂离子	lithium	lithium	lithium	graphene
6	lithium	graphene	材料	graphene	lithium
7	graphene	锂离子	锂离子	锂离子	锂离子
8	负极	battery	battery	负极	battery
9	battery	负极	石墨	battery	负极
10	石墨	石墨	负极	石墨	石墨

表 9　基于题名、摘要、关键词三者组合的检索词抽取方法对比

序号	MF	MF_TR	TR	TF-IDF	LDA
1	石墨烯	石墨烯	石墨烯	纳米	石墨烯
2	材料	材料	材料	复合材料	材料
3	性能	性能	性能	容量	性能
4	纳米	制备	制备	循环	纳米
5	电池	纳米	电池	mAh	制备
6	制备	电池	纳米	电池	电池
7	容量	结构	结构	Li	容量
8	结构	容量	复合材料	g-	结构
9	复合材料	复合材料	锂电池	Fe	复合材料
10	循环	循环	容量	结构	循环

观察表 5 至表 9 可知:

基于题名的方法对比: 每种方法的结果基本一致, 只有个别词的位置会有些许不同, 原因在于所检出文献的题名较少, 因而基于题名进行关键词抽取时各个方法的区分度不大。

基于摘要的方法对比: 每种方法抽取的关键词大体一致, 与检索词有关, 但是方法不同, 词语的排序也不太一致, 与基于题名的方法相比, 基于摘要的方法中, 词语有 70%与基于题名的方法是完全相同的。

基于关键词的方法对比: 基于检出文献的关键词提取检索词其总体效果比使用题名和摘要组合效果好, 其原因在于关键词是作者对文献内容反复考量后, 是作者从文献内或文献外选择出来用以表示全文主题内容的单词和术语, 因而其在主题、专业性方面具有天然的优势。

关键词反映的侧重点: 从题名和摘要二者提取的关键词更着重于检索词所表达的内在特征, 如复合材料、纳米等词, 而基于文献的关键词提取的关键词则更多反映其同位、上下位和相关关系, 如 lithium、graphene 等。

方法对比分析: MF_TR 实际上融合了传统的 TextRank 方法与 MF 方法, 从实验结果上可以看到, 由于将词频作为权重因子对 TR 进行改进, 因而其结果大体与 TR 一致, 但是受高频词的影响, 与高频词有共现关系的词其最终的排序得到提升。这也与笔者的初衷相吻合。文献检索的目的就是找出相关的检索词, 而检索式中的检索词与检出文献密切相关, 这些检索词出现的频次较高, 而充分挖掘与高频词有共现关系的其他词正是目的所在。

基于题名、摘要、关键词三者的组合进行关键词提取时其效果倾向于基于题名和摘要的组合效果, 其原因在于题名和摘要的组合的词语个数明显高于关键词的个数, 在进行提取时, 无论是按词频还是按词语共现关系, 其更多反映的是文字较多的摘要, 而关键词在检出文献中含量较少, 因而在融合的过程中被忽略。

③迭代抽取

为了验证候选检索词的迭代抽取效果是否能真正帮助查新人员快速找到相关的检索词, 以抽取的候选检索词与实际查新案例最后所使用的检索词进行对比。本案例中, 委托人提供的信息如图 4 所示:

> **题目:** 我国莱姆病的发现与研究
>
> **科学技术要点:**
>
> (1) 从 1988 到 1990 年在我国首先大规模地进行了莱姆病流行病学调查, 详细阐述了我国莱姆病疫源地特征;
>
> (2) 在我国首次证明全沟硬蜱可经卵传递莱姆病螺旋体;
>
> (3) 在国内首先应用先进技术分析了不同地理株莱姆病螺旋体蛋白、脂肪酸成分; 首先用电镜观察了不同地理株螺旋体的超微结构;
>
> (4) 在国内首先成功地制备了伯氏疏螺旋体的单克隆抗体, 并用于病原体鉴定和实验诊断。
>
> **查新点:**
>
> 查上述科学技术要点国内外是否有同类研究, 并对其新颖性作出判断。
>
> **最后的检索报告中包含的检索词包括:**
>
> 莱姆病; 流行病学; 全沟硬蜱; 伯氏疏螺旋体; 单克隆抗体

图 4　检索词的迭代抽取案例

第一次迭代:

1) 根据科技查新的题目自动生成检索式: “我国 and 莱姆病 and 发现”进行检索。检出文献共 85 条, 基于检出文献的关键词, 利用 MF_TR 算法, 返回的 10 个候选检索词包括: 莱姆病、螺旋体、试验、基因型、多态性、流行病学、血清、诊断、病毒、伯氏。

2) 术语扩展。根据公式(4)对候选关键词计算其领域重要性得分。

3) 领域重要性得分较高的前 10 个词为: 莱姆病螺旋体、**莱姆病**、**伯氏疏螺旋体**、**流行病学**、流行病学调查、螺旋体、间接免疫荧光试验、诊断、酶联免疫吸附试验、基因型。

第二次迭代:

1) 根据第一次迭代过程中产生的前 10 个检索词进行二次检索。通过选择"莱姆病螺旋体"作为检索式进行检索。检出文献共 130 条,基于检出文献的关键词根据 MF_TR 算法抽取,返回的 10 个候选检索词包括:莱姆病、螺旋体、蛋白、基因型、伯氏、表达、宿主、全沟、流行病学、传播。

2) 候选检索词扩展。与第一次迭代过程类似,扩展后得到:莱姆病螺旋体、**莱姆病**、**伯氏疏螺旋体**、**全沟硬蜱**、基因型、流行病学调查、限制性片段长度多态性、**流行病学**、实验经期传播、螺旋体。

合并:迭代过程产生的检索词进行合并,由于每次迭代过程产生的语料库不同,因而可能同一个检索词在不同的迭代过程中得分并不一致,此时取分值较高的值参与排序。

通过两次迭代检索并最终合并后,生成最终的检索词列表:莱姆病螺旋体、**莱姆病**、**伯氏疏螺旋体**、**流行病学**、流行病学调查、**全沟硬蜱**、螺旋体、基因型、间接免疫荧光试验、诊断。

通过与最终的检索报告对比,在合并得到的前 10 个检索词中,莱姆病、流行病学、伯氏疏螺旋体、全沟硬蜱这 4 个检索词完全匹配,召回率为 80%。

6　结论及展望

本文基于科技查新过程检出的实时相关语料为基础,提出一种检索词智能抽取方法,并以此为领域知识的来源,采用关键词抽取、领域特征扩展相结合的递进式迭代抽取方式对检索词抽取进行系统实现。通过与实际查新案例所使用的检索词比较发现,使用本方法两次迭代后抽取 10 个检索词,召回率达到 80%。

由于本研究通过网络爬虫在线获取文献集合,对候选检索词的获取来说,通常情况下,需要一定数量的科技文献,而这也意味着耗费的时间比较长,因而后续的研究将会在实际的科技查新业务中通过实践、摸索寻找一个平衡点。另外,候选检索词的抽取效果与文献数据库所收录论文的规范性有很大关系,尤其是拼写错误会对英文候选检索词的抽取产生较大的影响,因而在后续的工作中,对错误的自纠自查也是未来努力的方向。

参考文献:

[1] 黄江玲. 影响科技查新质量的重要因子分析[J]. 情报探索, 2008(8): 67-68. (Huang Jiangling. Analysis of Important Factors Affecting the Quality of Science and Technology Novelty Search [J]. Information Research, 2008(8): 67-68.)

[2] 曹欢增. 提高科技文献查全率的几项措施[J]. 科技情报开发与经济, 2008, 18(32): 72-74. (Cao Huanzeng. Some Measures for Increasing the Recall Ratio of Sci-tech Literatures [J]. Sci-Tech Information Development & Economy, 2008, 18(32): 72-74.)

[3] 陈予琳. 关键词检索方法在科技查新中的应用研究[J]. 河南师范大学学报: 自然科学版, 2011, 39(3): 171-173. (Chen Yulin. Keyword Search Method Application Research on Science and Technology Novelty Check [J]. Journal of Henan Normal University: Natural Science Edition, 2011, 39(3): 171-173.)

[4] 张柏秋, 吴晓鑛. 科技查新检索中的关键词选择[J]. 情报科学, 2008, 26(9): 1344-1348. (Zhang Baiqiu, Wu Xiaohuang. Keywords Selection in Science Technology Novelty Retrieval [J]. Information Science, 2008, 26(9): 1344-1348.)

[5] Hasan K, Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art [C]. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 1262-1273.

[6] Frank E, Paynter G W, Witten I H, et al. Domain-specific Learning Algorithms for Keyphrase Extraction [C]. In: Proceedings of the 16th International Conference on Artificial Intelligence (IJCAI-99), 1999: 668-673.

[7] Turney P D. Learning Algorithms for Keyphrase Extraction [J]. Information Retrieval, 2002, 2(4): 303-336.

[8] Nguyen T D, Kan M-Y. Keyphrase Extraction in Scientific Publications [C]. In: Proceedings of International Conference on Asian Digital Libraries (ICADL), 2007: 317-326.

[9] Lopez P, Romary L. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID[C]. In: Proceedings of International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010: 248-251.

[10] Krapivin M, Autayeu M, Marchese M, et al. Improving Machine Learning Approaches for Keyphrases Extraction from Scientific Documents with Natural Language Knowledge [C]. In: Proceedings of the Joint JCDL/ICADL' International Digital Libraries Conference, 2010: 102-111.

[11] Jiang X, Hu Y, Li H. A Ranking Approach to Keyphrase Extraction [C]. In: Proceedings of International ACM SIGIR

Conference on Research and Development in Information Retrieval, 2009: 756-757.

[12] Turney P D. Coherent Keyphrase Extraction via Web Mining[C]. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003: 434-439.

[13] Kumar N, Srinathan K. Automatic Keyphrase Extraction from Scientific Documents Using N-gram Filtration Technique [C]. In: Proceedings of the 8th ACM Symposium on Document Engineering. 2008: 199-208.

[14] 潘丽敏，吴军华，林萌，等. 融合多特征的中文关键词提取方法[J]. 信息网络安全, 2014(8): 40-44. (Pan Limin, Wu Junhua, Lin Meng, et al. Algorithm of Chinese Keywords Extraction Based on Multi-feature [J]. Netinfo Security, 2014(8): 40-44.)

[15] Hulth A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge [C]. In: Proceedings of Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2003: 216-223.

[16] Pasquier C. Task 5: Single Document Keyphrase Extraction Using Sentence Clustering and Latent Dirichlet Allocation [C]. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010: 154-157.

[17] 石晶，李万龙. 基于 LDA 模型的主题词抽取方法[J]. 计算机工程, 2010, 36(19): 81-83. (Shi Jing, Li Wanlong.Topic Words Extraction Method Based on LDA Model [J]. Computer Engineering, 2010, 36(19): 81-83.)

[18] 刘俊，邹东升，邢欣来，等. 基于主题特征的关键词抽取[J]. 计算机应用研究, 2012, 29(11): 4224-4227. (Liu Jun, Zou Dongsheng, Xing Xinlai, et al. Keyphrase Extraction Based on Topic Feature [J]. Application Research of Computers, 2012, 29(11): 4224-4227.)

[19] Mihalcea R, Tarau P. TextRank: Bringing Order into Texts [C]. In: Proceedings of EMNLP-04 and the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 404-411.

[20] Page L, Rrin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web [C]. In: Proceedings of the 7th International World Wide Web Conference. 1998: 1-14.

[21] 韩其琛，李冬梅. 基于叙词表的林业信息语义检索模型[J]. 计算机科学与探索, 2016, 10(1): 122-129. (Han Qichen, Li Dongmei. Semantic Model with Thesaurus for Forestry Information Retrieval [J]. Journal of Frontiers of Computer Science & Technology, 2016, 10(1): 122-129.)

[22] 熊霞. 基于叙词表词间关系的领域信息检索[D]. 北京：中国农业科学院, 2011. (Xiong Xia. Domain Information Retrieval Based on Term Relationships of Thesaurus [D]. Beijing: Chinese Academy of Agricultural Sciences, 2011.)

[23] Hulth A, Karlgren J, Jonsson A, et al. Automatic Keyword Extraction Using Domain Knowledge [C]. In: Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics, 2001: 472-482.

[24] Coursey K H, Mihalcea R, Moen W E. Automatic Keyword Extraction for Learning Object Repositories [J]. Proceedings of the American Society for Information Science & Technology, 2009, 45(1): 1-10.

[25] Li G, Wang H. Improved Automatic Keyword Extraction Based on TextRank Using Domain Knowledge [C]. In: Proceedings of the 3rd CCF Conference, NLPCC 2014, Shenzhen, China. 2014, 496: 403-413.

[26] Jiang B, Xun E, Qi J. A Domain Independent Approach for Extracting Terms from Research Papers[C]. In: Proceedings of the Australasian Database Conference. Springer International Publishing, 2015: 155-166.

[27] Lopes L, Fernandes P, Vieira R. Estimating Term Domain Relevance Through Term Frequency, Disjoint Corpora Frequency-TF-DCF [J]. Knowledge-Based Systems, 2016, 97: 237-249.

[28] 詹恒飞，杨岳湘，方宏. Nutch 分布式网络爬虫研究与优化[J]. 计算机科学与探索, 2011, 5(1): 68-74. (Zhan Hengfei, Yang Yuexiang, Fang Hong. Research and Optimization of Nutch Distributed Crawler [J]. Journal of Frontiers of Computer Science & Technology, 2011, 5(1): 68-74.)

[29] 卢萍，蔡群. 中文科技论文关键词的标引[J]. 广州医学院学报, 2000, 28(2): 93-94. (Lu Ping, Cai Qun. Keyword Indexing of Chinese Scientific and Technical Paper [J]. Academic Journal of Guangzhou Medical College, 2000, 28(2): 93-94.)

[30] Guo C, Lu X. Selecting Publication Keywords for Domain Analysis in Bibliometrics: A Comparison of Three Methods

[J]. Journal of Informetrics, 2016, 10(1): 212-223.

[31] 洪道广. Google Scholar 的数据整合研究[J]. 现代情报, 2010, 30(7): 39-41. (Hong Daoguang. Research on Data Integration of Google Scholar [J]. Modern Information, 2010, 30(7): 39-41.)

[32] Rossi R G, Maracini R M, Rezende S O. Analysis of Domain Independent Statistical Keyword Extraction Methods for Incremental Clustering [J]. Learning and Nonlinear Models, 2014, 12(1): 17-37.

作者贡献声明:

王培霞: 提出研究思路, 设计研究方案, 进行实验, 撰写论文;

余海, 陈力: 修改论文;

王永吉: 研究方案修改, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: peixia@nfs.iscas.ac.cn。

[1] 王培霞. 实验例子.csv. 文献数据信息.

收稿日期: 2016-07-28
收修改稿日期: 2016-09-26

Using Intelligent System to Extract Search Terms for Sci-Tech Novelty Retrieval

Wang Peixia[1,2] Yu Hai[1,2] Chen Li[1,2] Wang Yongji[1]

[1](Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper aims to identifying the search terms more effectively in sci-tech novelty retrieval, which could reduce the subjectivity, heavy workload, de-normalization and time-consuming issues facing the manual methods. [**Context**] We used the corpus generated by the sci-tech novelty retrieval as the source of domain knowledge to extract search terms. Then, we discussed the relationship between the corpus and the keyword extraction. [**Methods**] We proposed an incremental iterative method to extract keywords from the sci-tech novelty retrieval project with the help of domain feature expansion. [**Results**] Compared to search terms from the real world sci-tech novelty retrieval, the recall rates of the 10 search terms extracted by the new method reached 80%. [**Conclusions**] The proposed method could identify most keywords and then improve the efficiency and effectiveness of the novelty retrieval tasks.

Keywords: Sci-Tech novelty retrieval Search terms Keywords extraction Online crawler